Celebrating the history of chemical information

RSC Chemical Information and Computer Applications Group (CICAG) and the RSC Historical Group in conjunction with the CSA Trust

Monday 29th November 2010

A summary by Wendy A. Warr

This one-day meeting was ably organised by Bill Griffiths and Chris Cooksey of the RSC History Group, Doug Veal of RSC CICAG, and Diana Leitch, who belongs to both groups, and who chaired the event on the day. Peter Rhodes and Doug Veal organised an intriguing exhibition of photographs, articles and chemical information artifacts. Despite highly unfavourable weather conditions, and a strike on London Underground, over 100 people attended including nine bursary students from Sheffield, Oxford and Cambridge. A highlight of the programme was the presentation of the UKeiG Tony Kent Strix Award for 2010 (http://www.ukeig.org.uk/awards/tony-kent-strix) to Mike Lynch in recognition of his theoretical work on information retrieval and his work leading to the development of the CAS registry service.

The proceedings began, appropriately, with a talk on the contribution of the RSC, given by Doug Veal. Way back in 1841, when the Chemical Society was founded, its objectives included "communication and discussion of discoveries and observations, an account of which shall be published in the form of proceedings or transactions" and "the formation of a library of scientific works". Records show that the first bookcase was purchased in 1842. We had lunch in the current library, now part of the newly refurbished Chemistry Centre. Like many other libraries, it has morphed into a virtual library and information centre, but there are still plenty of books on view in today's bookcases.

The RSC (http://www.rsc.org) was formed in 1980 from the Chemical Society and other like organisations; the chemical information group was already in existence in 1967. Doug described the contribution of the RSC to chemical information in seven areas: primary publications, book publishing, secondary publications and services, library and information services, special interest groups, international collaboration, and recent developments. Few of us will forget his example of the snags of truncation in Keyword in Context (KWIC) indexes in the 1970s: "atherosclerotic tendencies of Trappist monks" without its first nine letters represents the monks' activities in a new light!

Wearing my hat as a learned ACS editor, I do have to raise an eyebrow at Doug's presentation of the Impact Factors of RSC journals. ACS has 29 journals with an Impact Factor over 3.0; RSC has only 17. If you use *percentage* of journals with an IF of more than 3, it does indeed seem as if RSC journals have a very slightly higher impact, but by all other metrics ACS wins. Putting such quibbling aside, I do applaud Doug's final slides: no one should criticise RSC's recent, truly innovative activities: the award-winning RSC Prospect (http://www.rsc.org/Publishing/Journals/ProjectProspect/Examples.asp) and the ChemSpider community on the Internet (http://www.chemspider.com), with more than 25 million searchable structures.

Next up, Bill Town talked about the language and symbolism of chemistry. He said that covering 2000 years of history in 30 minutes wasn't going to be easy; it is even harder to condense it into 300 words! He started with the huge debt we owe to Arabic scholars and followed that with some fascinating material on the alchemist's view of chemistry. He then jumped to the 18th century. Chemical nomenclature was confusing until the late 18th century, but from 1787 things started to improve in France following the efforts of Guyton de Morveau, Lavoisier, Berthollet, and Fourcroy. Symbols were also confusing, and duplicated, and required special typefaces. A breakthrough came in 1803, with John Dalton's atomic theory, and things improved further from 1813 when Berzelius introduced his chemical symbols.

The first edition of the Gmelin Handbook appeared in 1817-19, and by the fourth edition included some organic compounds, but organic chemistry was still in its infancy. A defining moment was the Karlsruhe Congress, of 1860 which sought to define chemical ideas such as "atom", "molecule", "equivalent", "atomic", and "basic". It also examined the question of equivalents and chemical formulae (Wikipedia tells me that nineteen different formulae were used by chemists for acetic acid at that time) and tried to establish a uniform notation and nomenclature. A breakthrough in organic nomenclature came in 1861, when Butlerow formally introduced the concept of structure, allowing only one "rational formula" for each compound. Beilstein's *Handbuch der organischen Chemie* was first published in 1881. The Geneva Congresses of 1889 and 1892 produced a reformed nomenclature based on the principle of substitution and the use of molecular formulae.

The discovery of the periodic table of the elements is usually attributed to Dmitri Mendele'ev in 1869, but John Newlands, Lothar Meyer and Alexandre-Émile Béguyer de Chancourtois independently also contributed to the recognition of the periodicity of the elements. The RSC visual elements periodic table (http://www.rsc.org/chemsoc/visualelements/pages/periodic_table.html) is now available as a jigsaw puzzle, a wall chart, a T shirt, and a card game, amongst other things. Bill ended his talk appropriately, to the amusement of the audience, with an animated version of Tom Lehrer's *The Elements Song*.

There was no mention of IUPAC in Bill's talk because he had agreed to stop by the end of the 19th century, but it has been argued that the Karlsruhe meeting was the first international meeting of chemists, and that it led to the eventual founding of IUPAC, though some might think that the 1911 Conseil Solvay was the spur. As a concluding note, it is worth mentioning that IUPAC is behind the International Year of Chemistry (http://www.chemistry2011.org/) which we will all celebrate in 2011.

Now let us return to history and Engelbert Zass's talk on the chemical literature. The roles of the literature are disclosure of scientific results, public discussion of research results, critical comparison of one's own work with that of others, justification of financing, and scientific recognition. In the early days chemists wrote letters to each other and had meetings in their academies; the proceedings of these meetings turned into journals, the earliest of which were *Journal des Sçavans* and *Philosophical Transactions of the Royal Society of London* in 1665. Journals make up a major part of the primary literature; CAS covers about 20,000 serials. The secondary literature consists of abstracting and indexing publications, now databases such as *Chemical Abstracts* and *Chemisches Zentralblatt* (begun as *Pharmaceutisches Central-Blatt* in 1830 and now available electronically

http://infochem.de/products/databases/czb.shtml) and Handbooks such as those of Gmelin (begun in 1817), and Beilstein.

Bert outlined the history of *Chemical Abstracts* from 1907 to 2009. The first version of CAS REGISTRY could not handle stereochemistry; salts were also a problem and the dot disconnect molecular formula still mystifies today's students. Searching the Gmelin and Beilstein Handbooks used to require knowledge of complex classification systems, and of German. Both Handbooks were later transformed by CrossFire, but Gmelin structures for many simple compounds were missing. That problem was solved by Reaxys: nowadays both Handbooks can be searched together in Reaxys, but some Gmelin data still need to be added.

Friedrich Konrad Beilstein was an assistant of Carl Löwig, who produced a multi-volume *Chemie der Organischen Verbindiungen* in 1840. The feasibility study into producing an electronic version of *Beilstein* began in 1982. Sandy Lawson produced SANDRA to help users cope with the classification system. *Beilstein* was mounted on STN International in 1988; Beilstein Current Facts appeared in 1991; CrossFire Beilstein started in 1993; and Beilstein was absorbed into Reaxys in 2009. Incidentally, I personally classify Handbooks as the tertiary literature; Bert presented Kissmann and Wexler's definition: "Original research is compiled, reorganised, reformatted and presented as a coherent whole rather than as a series of isolated reports". Ullmann's Encyclopedia is an example.

Some landmarks are CAS' Chemical Titles (1961), CA Condensates (1968), DARC and CAS Online (1980/1981), MDL's REACCS (1982), CAS e-journals (1986) the World Wide Web (publicly available from 1991) and CrossFire (1993). In 1970 chemical information consisted of isolated print sources for use only by information specialists. By 1985 isolated electronic sources were available for end users, and by 2000 integrated electronic sources for end user chemists were commonplace. Bert showed some slides of how his own library has changed to take advantage of RSS feeds, e-journals, e-books, and full text patents online. In physics and biology, preprints and databases are freely available, but chemistry has lagged behind, despite the fact that chemistry was originally a ground-breaker because its language was the chemical structure.

Chemical structures were the theme of Phil McHale's talk. I think that most of the audience will agree that this presentation was the most enjoyable of the day. There is no way I will be able to convey Phil's humour and dynamism, and the slides themselves are almost all factual, so, with apologies, here are just the facts. Molecular composition, with atoms (and valence, charges, radicals) and bond types, can be represented by a molecular formula (e.g., $C_4H_8O_2$), linear formula ($CH_3.CO_2Et$), systematic nomenclature (ethyl acetate), and structural formula. Drawing conventions evolved to match chemical understanding; IUPAC has published recommendations (Graphical Representation Standards for Chemical Structure Diagrams. *Pure Appl. Chem.* **2008**, *80*(2), 277-410). The basic needs of a structure-based system can be summarised in the mnemonic "RSVP" for "Register, Search, View, and Print/Publish".

Chemical typewriters (costing $18,000 in 1966!) gave way to graphics terminals and then PCs with graphics. Connection tables for 2D and 3D structures have largely replaced coding systems such as fragments and linear notations in RSVP systems (NIH/EPA, CAS, CROSSBOW, MACCS, DARC, OSAC, etc.),

although fragments still have their uses, the SMILES notation is still commonplace, and a new "notation" InChI (http://www.iupac.org/inchi/) is being widely used for exchange of chemical information. Some of the problem areas that the Wiswesser Line Notation (WLN) faced in the 1970s (mixtures, alternatives tautomers, polymers) are still unsolved in the brave new world of InChI. Combinatorial libraries, generic structures, reaction mechanisms, transition states, and animation have unfinished business even now. "Ferrobscene" is represented in multiple ways in online databases and even the ChemSpider version is "not acceptable" by IUPAC guidelines.

Connection tables and graphics-based systems led to "disintermediation": they gave chemical structures back to the practising chemist. They have enabled more than RSVP. Structures can now be used to calculate and predict properties, spectra and names; enrich publications (as in RSC Prospect); index databases; link datasets; carry out retrosynthetic analysis and quantitative structure activity relationships (QSAR); study "drugability"; and perform clustering and diversity analysis. 3D structures can be generated from 2D ones and used to explore ligands and receptors in drug discovery. Finally Phil brought the topic right up to date by giving some examples of mobile chemistry.

Phil's talk fitted nicely before Helen Cooke's one on chemical databases, since she started with the truism that the structure diagram has been the preferred communication language for chemists since the early 1900s. Most of the questions chemists ask are focused on compounds (their structures, reactions, and properties). Publishers recognised this and sought ways to provide structure-focused searching, through innovative classification schemes (as, for example in *Beilstein*), indexing in accordance with IUPAC or CAS nomenclature standards, and standards such as the Hill system for ordering chemical formulae.

Chemical database producers sought to replicate and improve on printed secondary information sources. Punched and edge-notched cards were the technologies of the mid-1950s. KWIC systems were used for CAS's Chemical Titles in 1961. The CAS REGISTRY system was initiated in 1964. Drivers and enablers behind chemical databases were advances in IT, telecoms, and networks; an explosion in the chemical literature; commercial impetus from the pharmaceutical and chemical industries; the development of chemical codes, languages and notations; and research breakthroughs such as those at Sheffield University.

Helen did not mention this, but one of the drivers of online searching in the 1970s (e.g., Lockheed Dialog) was spare computer power from the drive to put a man on the moon. In-house chemical structure search systems (many using WLN) were in place before much technology was available for structure searching of public databases. A Chemical Notation Association meeting at Daresbury in March 1980 was a landmark in the early days of chemical structure searching of the literature; the demand from industry was loud and clear and the money was available.

Helen's landmarks in chemical database history were as follows:

1964 CAS REGISTRY system
1965 Cambridge Structural Database
1972 ISI Science Citation Index online

1976 Derwent World Patents Index online
1978 Molecular Design Limited founded
1980 CAS ONLINE
1984 CAS ONLINE incorporated into STN International
1988 Beilstein Online on STN
1989 Beilstein Online on Dialog
1991 Gmelin Online
1993 Beilstein CrossFire
1995 CAS SciFinder.

Many structure searchable databases are now available. The roles of chemists and information professionals have changed during the era of databases, with the pendulum having now swung back to empowered end user chemists after a period when the key databases were available on a pay-as-you-go basis only, which tended to restrict access to information professionals.

Sandy Lawson's talk was more concerned with data than with chemical structures *per se*. He started by asserting that, despite the concept of the scientific method, all data are *not* created for a hypothesis and not all are worth preserving. All data have their hour in the limelight. For example, in the 18th century gravimetric data were key. Since 1980 there has been increased interest in ecological and biological data and in reaction refinement (e.g., regioselectivity).

In outlining some history of hypotheses and models, Sandy pointed out that the phlogiston theory could be true if a phlogiston atom had an atomic weight of -16. The theory died in 1775 but vitalism only retreated in 1828 when Wohler synthesised urea from ammonium cyanate (an inorganic substance). It could be that the phlogiston theory died sooner than vitalism because chemistry is based on concepts that are hard to express numerically. The work of Mendele'ev was a landmark in the numerical basis of chemistry. Referring to Bill's talk, Sandy commented that Mendele'ev *predicted* while Newlands did not. Traditionally, Sandy sees the CAS database as a comprehensive bibliographic database of *concepts* and Beilstein CrossFire as a comprehensive *factual* database.

For over 100 years, and into the 20th century, synthesis from a known material was the ultimate method for proof of structure. That era came to an end with tools such as NMR and X-ray structure determination. In the 20th century property space was explored. The Woodward-Hoffmann Rules or orbital symmetry proved much more powerful for predicting reactivity than using structures (such as Phil's "ferrobscene") that are limited by valence bond representations. In the 21st century we have a better understanding of reaction catalysis.

In summary, Sandy said that experimental data are the parents of workable models, and ultimately, knowledge. The most important data are probably those which do not fit the current model, provided they are properly measured. Some experimental data must always be available for checking (for validation), but they have a shelf life in the shop windows of the secondary information services. Much of the historical background in Sandy's presentation was found on the Internet, but he hesitated to say that he had acquired knowledge; instead he was capable of finding considered, evaluated responses to his queries. That may well be a model for consideration in the major databases.

The next speaker was Peter Willett who entitled his talk "Chemoinformatics: historical development of database methods". Chemoinformatics (or "cheminformatics" as it is more commonly called) emerged as a discipline (http://warr.com/warrzone2000.html) in the late 1990s, spurred on by the data explosion resulting from the introduction of combinatorial chemistry, but its roots go back much further. The core journal, the *Journal of Chemical Documentation* (later *Journal of Chemical Information and Computer Sciences* and now *Journal of Chemical Information and Modeling*) is celebrating its 50th anniversary in 2010.

Peter selected a number of seminal papers, the first of which reported a graph matching algorithm (Ray, L. C.; Kirsch, R. A. *Science*, **1957**, *126*, 814-819) which could be used to search molecules represented as graphs. The Morgan algorithm (*J. Chem. Doc.* **1965**, *5*, 107-113), an approach for the naming of chemical graphs, is fundamental to most registry systems (including that of CAS). Substructure searching uses a subgraph isomorphism algorithm, made practical by application of an initial screening procedure based on chemical fragments. The fundamental research into an appropriate choice of fragments was done at Sheffield University in the early 1970s (Crowe J. E. *et al*. 1970 *J. Chem. Soc. (C)* 1970, 990-996; Bill Town was a co-author).

Computer-aided synthesis design systems, which derive potential syntheses of a target molecule, were another early development. One of them, LHASA, is still in use today (Corey, E. J.; Wipke, W. T. *Science* 1969, 166, 178-193). Reaction databases can be traced back to an idea introduced by George Vledutz (*Information Storage and Retrieval*, **1963**, *1*, 17-146). This led, in the 1980s, to reaction searching systems based on the indexing of reaction centres: the parts of a molecule that change in a reaction.

Other new developments in the early 1980s were 2D similarity searching (an inspiration for which was published in Adamson, G. W.; Bush, J. A. *Information Storage and Retrieval* **1973***, 9,* 561-568) and searching of so-called Markush structures: the generic structures that occur in patents. Mike Lynch's team at Sheffield published a long series of papers on searching generic structures, leading to the operational systems currently offered by Thomson Reuters/Questel and CAS.

In the mid-1980s there was intense interest in 3D substructure searching and pharmacophores. According to IUPAC, a pharmacophore is an ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target and to trigger (or block) its biological response. A method developed at Sheffield was the basis of the first 3D systems at Pfizer and Lederle (Jakes, S. E.; Willett, P. *J. Mol. Graphics* **1986**, *4*, 12-20).

In the 1990s much work was done in the field of ligand docking (fitting a molecule into a binding site). A seminal publication, on the DOCK program, was published much earlier (Kuntz, I. D. *et al. J. Mol. Biol.* **1982**, *161*, 269-288). Technological developments meant that many more compounds could be made and there was a need for tools to quantify diversity and to select molecules so as to maximise diversity. Peter selected Brown, R. D.; Martin, Y. C. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572-584 out of a huge range of papers. Diversity alone, however, was not enough; it was necessary to optimise absorption, distribution, metabolism, and excretion (ADME) properties (Lipinski, C. A. *et al. Adv. Drug Del. Rev.* **1997**, *23*, 3-25). Lipinski's paper heralded the development of techniques for the quantification of "drug-

likeness". In short, chemoinformatics is not new; what *is* new is the widespread recognition of its importance, and this will increase further given the current challenges facing the pharmaceutical industry.

Bobby Glen continued that theme with his look to the future. One thing is certain; the scale of chemical data will continue to increase enormously. Already 180,079,582,087 DNA bases have been sequenced, 60 million chemical substances are known, and 50 million pages of European patents have been published. If only we could connect all the data. Bobby thinks that chemical information is all about networks. There is an awful lot of it, and it's embedded in a complex network of journals, articles, blogs, books, and people's heads. Unfortunately, the tools to find the "information" are inadequate.

Rosvall and Bergstrom have traced science networks (http://www.tp.umu.se/~rosvall/livemod.html) based on journal citations (Rosvall, M.; Bergstrom C. T. *Proc. Nat. Acad. Sci.* **2010**, *105*(4), 1118-1123). By analysing such networks, we can see patterns in "scientific thought". Suppose we are looking for a piece of information. The connection from our starting point to the information is often a "random walk". Can we quantify this and optimise the process? Finding the best route through a knowledge network is by choosing the best exit routes. This could lead to autonomous agents to search for personalised information.

An example of an autonomous agent is currently being developed by IBM: a negotiator which will negotiate with other robots for the best deal in purchasing electronics. Could we do the same for chemical information, for example design a synthesis? Chem4Word (http://chem4word.codeplex.com/) is being developed to be semantically rich; it can identify what you "intend" through chemical "intuition" and avoid errors. Could this be extended to discover synthetic reactions or even to suggest research topics?

More and more information has become publicly available recently, especially in bioinformatics, but the open science trend is spreading to chemistry with sources such as DrugBank, PubChem, ChemSpider and chEMBL. The sociology of chemical information has been transformed by the Internet. We have the data; there is now an opportunity for the computer to help humans find knowledge. Developments in natural language processing and semantic analysis of documents, as in RSC Prospect, could eventually go beyond the processing of legacy data and lead to new authoring tools.

Earlier talks had covered the representation of chemicals but Bobby pointed out that real chemicals don't exist as connection tables. The next generation of chemical information tools should capture the history of the materials which went to make up the substance, as well as measured and predicted properties, including human and robotic senses, video, scene interpretation and spectroscopy. To do this, we will need a sort of robot chemist to help capture data.

Data capture will be part of "pervasive computing". Bobby's team is working on an electronic laboratory notebook which could be extended to a virtual laboratory notebook with intelligent fume cupboards, safety monitoring, reaction monitoring, and control. This is a sort of Second Life for chemists. Information could be accessed not just from today's smart phones but also from more futuristic devices. Chemical information will increasingly be derived from simulation: simulation of what is in a flask not

just simulation of a molecule. Bobby showed examples of DNA sensors and coarse-grained simulations of membrane proteins (work being carried out by Peter Bond in Bobby's team at Cambridge).

Bobby concluded with some more down to earth conclusions. Quality will become even more important lest we become engulfed in a sea of "bad science". The library will become the Internet and *vice versa*. Librarians will become ever more expert at search, cataloguing and technology access, and identifying trusted sources of data. Technology will be driven by social computing, gaming and commerce and will continue to accelerate. Competition to established publishers will appear from unexpected directions: crowdsourcing, peer-to-peer models of data, open repositories and so on. Bobby's final slide makes an appropriate ending for this short report: "Everything is chemistry; chemistry is everything".